

Statistical Analysis of Medicare Payments

ISM 6137 Statistical Data Mining

Prof. Anol Bhattacharjee

Muma College Of Business

University of South Florida

Shreyas Thombare

Ninad Mehta

Kaushik Kasthurirangan



Table of Contents

1. Executive Summary	1
2. Problem Definition/Significance	4
3. Prior Literature	5
4. Data Source Preparation	6
5. Exploratory Data Analysis	7
6. Data Preprocessing/Feature Engineering	9
7. Variable Choice	11
8. Models and Interpretations	13
9. Recommendations and Future Scope	18
10. References	19
11. Appendix	20



1. Executive Summary

1.1 Medicare Payments

Medicare is a federal government health insurance plan that was initialized in 1966 under the Social Security Administration (SSA). Presently, it is being governed by the Centers for Medicare and Medicaid Services. Medicare provides health insurance plans for people aged 65 and older. It also provides special plans for younger people with disabilities. Moreover, this plan includes coverage for end stage renal diseases and amyotrophic lateral sclerosis (ALS).

In 2017, there were 58 million individuals who were provided benefits under medicare across the United States. According to annual Medicare Trustees reports and research by the government's MedPAC group, Medicare covers about healthcare expenses for at least half of those enrolled. There are four different parts to Medicare, each of which has been explained in detail below,

Part A: Covers the expenses for formally admitting a patient to any hospital or nursing home (only after being formally admitted to a hospital for three days and not for custodial care), and hospice services.

Part B: Covers specific doctors' services, outpatient hospital costs and medical supplies.

Part C: Part C is an alternative known as Managed Medicare or Medicare Advantage which allows patients to choose health plans with at least the same service coverage as Parts A and B (and most often more), often the benefits of Part D.

Part D: It covers the cost of prescription drugs such as vaccines or recommended shots.

1.2 Inpatient Prospective Payment System (IPPS)

The payment made by medicare for hospital inpatient stays is based on preset rates under Part A. This payment system is referred to as the inpatient prospective payment system (IPPS). Under the IPPS, every patient's medical condition is grouped into a diagnosis-related group (DRG). Each DRG has a payment weight assigned to it, based on the average resources used to treat Medicare patients in that DRG. The base payment rate is divided into two parts of labor and non labour payments. Labor-related share is adjusted based on where the hospital is located. the non labor share is adjusted by a cost of living adjustment factor. This base payment rate is multiplied by the DRG relative weight.

1.3 Total Performance Score (TPS) of Hospitals

Total performance of any hospital is a weighted sum of scores from four domains,

- (1) Clinical Outcomes Score
- (2) Efficiency and Cost Reduction Score
- (3) Safety Score
- (4) Community and Engagement Score.

- **Clinical outcomes score** for each hospital is assessed based on the 30-day mortality rate for Acute Myocardial Infarction, Heart Failure and Pneumonia.
- **Efficiency and Cost Reduction scores** are awarded based on the medicare spend per beneficiary for the particular hospital.
- **Safety scores** are a function of various infection rates such as Catheter-Associated Urinary Tract Infection, Central Line-Associated Bloodstream Infection, Surgical Site Infection, Clostridium difficile infection and so on.
- **The person and community engagement score** is composed of 8 dimensions derived from the HCAHPS Survey namely Communication with Nurses, Communication with Doctors, Responsiveness of Hospital Staff, Communication about Medicines, Discharge Information, Hospital Cleanliness & Quietness, 3-Item Care Transition and Overall Rating of the Hospital .

1.4 Payment Reduction Percentages

The national average for Total Performance Scores of hospitals was 61.0 for the year of 2017. Medicare penalizes hospitals if their total performance score falls below the threshold of 59 by imposing a payment reduction during the settlement of claims.

Performance Score	Payment Reduction in %	
>= 59	No Penalty	- 0.5% payment penalty imposed for every 10 points of lower payment scores as opposed to the national avg. - Below a TPS of 29, a flat 2% of all medicare payments are withheld from the hospital.
49 - 58	0.5	
39 - 48	1	
29 - 38	1.5	
<= 28	2	



2. Problem Definition & Significance

With the introduction of the Inpatient Prospective Payment System (IPPS), it is assumed that all DRGs are paid out uniformly to the various hospitals with an adjustment for location and prevailing wage index. However, Medicare costs have been rising year on year.

While this increase is in part due to the increasing costs of drugs and its monopolization by pharma cos, hospitals are also constantly investing money in improving their facilities. With the hospitals seeking to make a profit on their investments, the concern over hospital overbilling Medicare arises.

Our analysis aims to answer the following questions,

1. Are medical providers paid the same across the United States by Medicare for each Diagnosis Related Group (DRG)?
2. How do extra charges (service charges) vary across different medical facilities in the state of Florida?
3. How do out-of-pocket charges vary across the United State for each Diagnosis Related Group (DRG)?

We have restricted our analysis to using Medicare IPPS data from 2017 for all DRGs, across the various hospitals in the



3. Prior Literature


In order to better understand the importance of the various parameters pertaining to our analysis, we reviewed previous literature on the regression analysis for adjustment of medicare payments.

Steven H. Sheingold in his paper titled “Using Multivariate Regression to Adjust Prospective Payment Rates” discusses potential adjustments for the payment rates across DRGs based on a few additional factors apart from the adjustments for location and prevailing wage index.

It is possible that because of their location, certain hospitals may receive a disproportionately high share of low-income patients. Some hospitals on the other hand may be teaching hospitals. For these reasons, he proposes adjusting payment rates based on measures such as Medicare Wage Index (MWI) and Case Mix Index (CWI). In addition to this, instead of adjusting based on location and population, he proposed taking into account a proportion derived from the population and MSA for a hospital’s location.

Adding on to Steven’s findings, “The Use of Regression Analysis to Determine Hospital Payment: The Case of Medicare’s Indirect Teaching Adjustment” proposes adjustments between teaching and non-teaching hospitals. In Kenneth E. Thorpe’s analysis, location once again played a major role with one of his key findings from the analysis being that patients on the East Coast had a 40% higher stay rate during inpatient treatments as opposed to those in the West Coast.

In addition to this, he proposed including Average Nursing Wage per FTE and Average Other Employee Wage per FTE as factors for adjusting the predefined payment rates to hospitals.



4. Data Source and Preparation

Our primary data source for the analysis was the Data portal on the Centers for Medicare & Medicaid Services' website. The Inpatient Prospective Payment System data file for 2017 provides a provider-level summary for the top 100 Diagnosis Related Groups. Each record represents an aggregated measure of medicare payments, provider billing and total covered charges for every combination of DRG & hospitals participating in the program.

Each medical provider was uniquely identified using the Provider ID a.k.a CMS Certification Number (CCN).

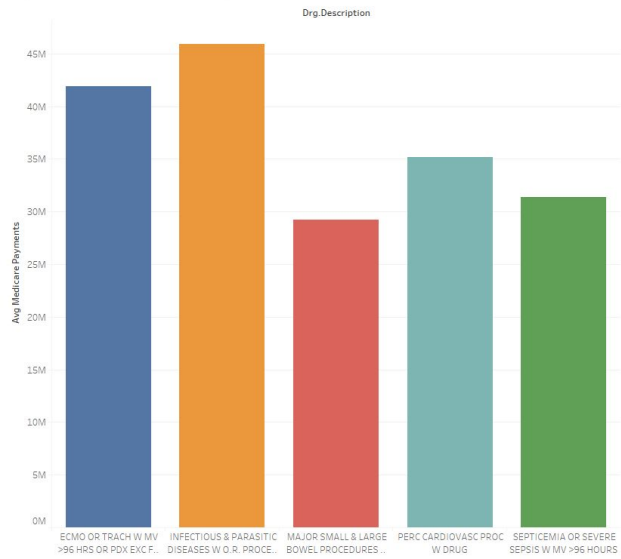
In order to control for the impact of the hospital's performance measures on the medicare payouts, we integrated the Hospital Value Based Purchasing System (HVBPS) data for the year 2017. Each record represents a participating medical provider's score summary across the four domains - safety, community & engagement, clinical outcomes and efficiency.

Our HVBPS data for providers was integrated into the IPPS data using the Provider ID as the common key.

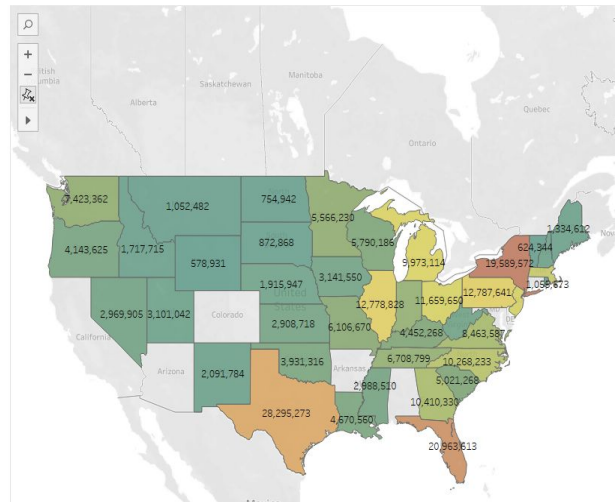
Finally, we used the Census Bureau's 2017 estimates for population across the various counties, cities and states and prevailing wage estimates to control for the differences in population and cost of living during our analysis.

5. Exploratory Data Analysis

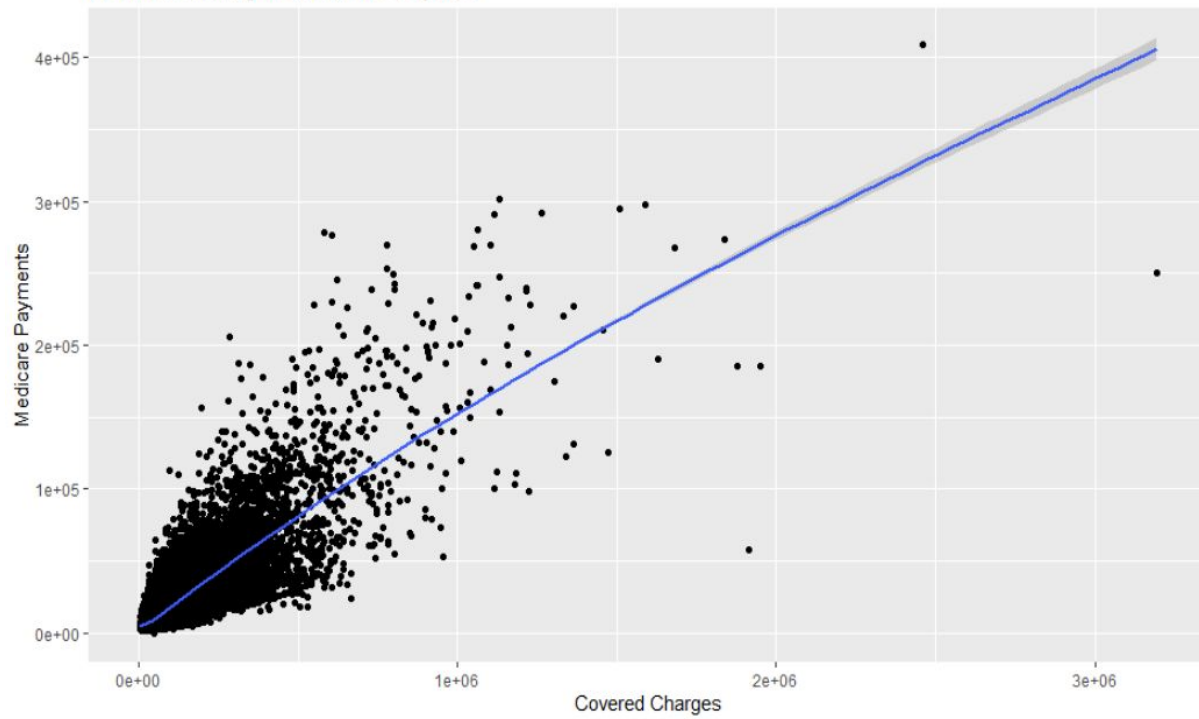
Top 5 DRG Codes - Medicare Payments

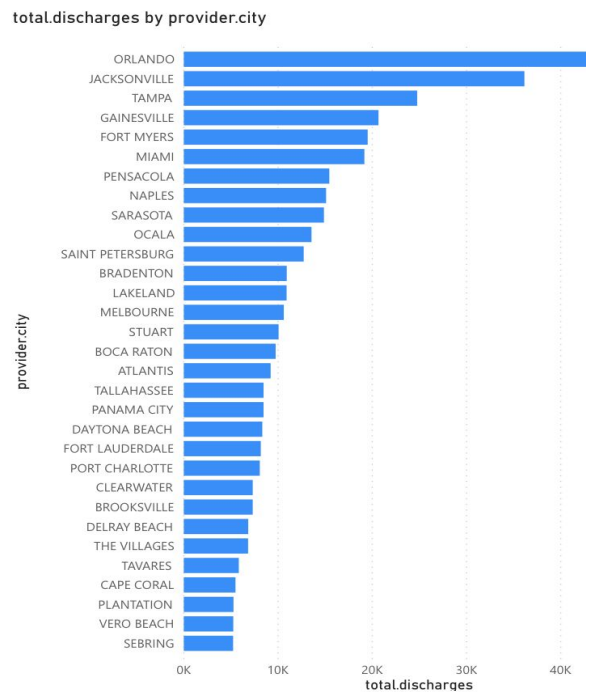
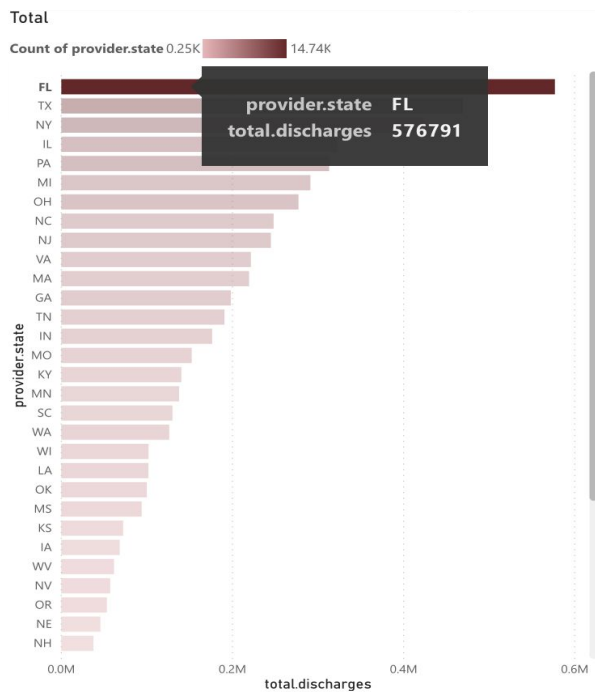
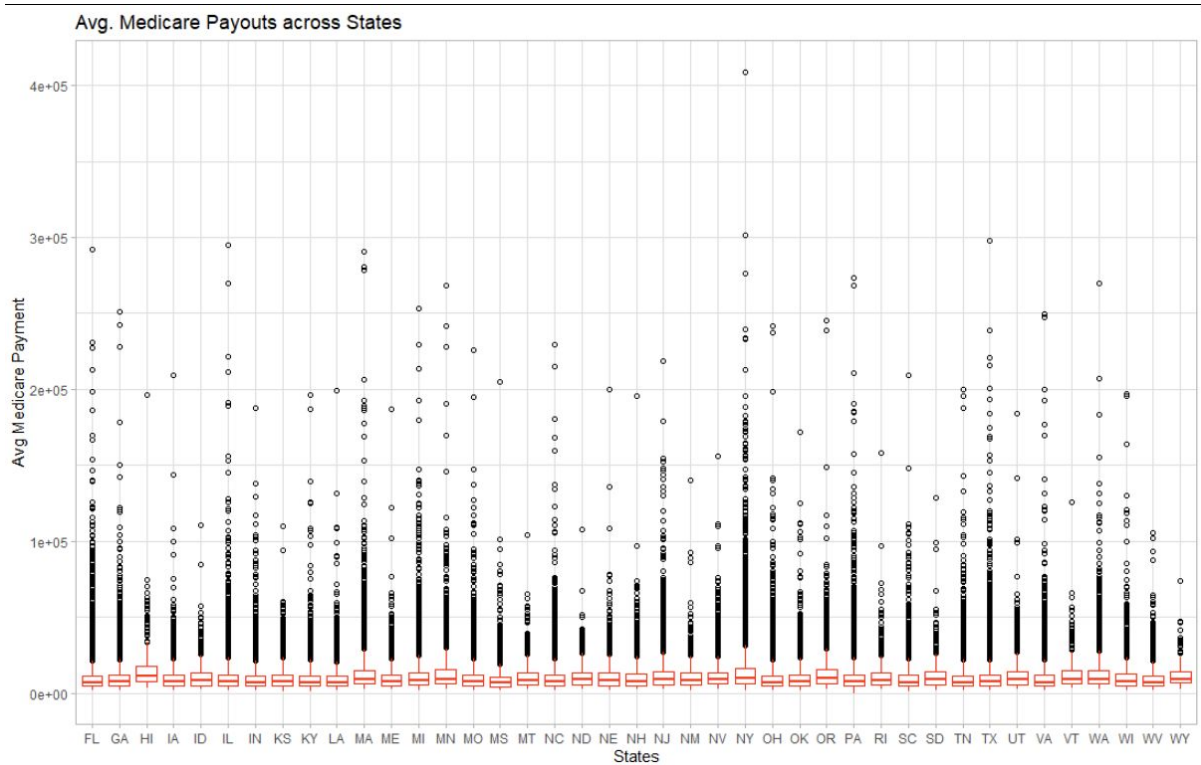


Avg Medicare Payment by State



Provider's Billing vs Medicare Payouts





6. Data Pre-processing and Feature Engineering

6.1 Data Preprocessing

In this project our primary audiences are the Medicare Officials who can use this analysis and interpretations to gain an insight into average differences between the medicare payments across all the states in the United States. Our analysis enables the audience to track average medical payments varying across DRG codes by state. Moreover, our analysis could be useful for insurance providers to analyse which DRG codes have most out of pocket expenses and extra payments as per providers. Our secondary audience are the patients who can use the recommendations to comprehend which facility is the most expensive in a particular region for a specific DRG Code.

Keeping this in mind we planned to generate relevant features and pre-process our data. We included statewise average household income and average population to control the effects in a more effective manner. With this Action Plan in mind we worked on the following steps for merging and preprocessing the data.

Data Merging

We merged IPPS, HVP performance, Average Income and Average Population dataset by performing an inner join on facility ID and state.

Data Transformation


- Converted column names to lowercase and renamed variables to achieve consistency. Splitted DRG definition into DRG code and DRG description.
- Verified if any features have missing values.
- Replaced missing values from performance variables with 'Not Available'
- Converted continuous variables to numeric and categorical variables into factors.
- Scaled predictors for standardization and rounded fraction up to 4 decimal places.
- Performed Log Transformation for variables with right skewed distribution.

6.2 Feature Engineering

Exploratory Data Analysis enabled us to understand the relevance of the features in our dataset. Moreover, after reading prior literature work we shortlisted some important predictors keeping in mind the business logic and technical significance. Our main goal of performing feature engineering was to reduce the number of features, remove correlated variables, and generate calculated features using the existing features.

After comprehending trends in the medicare payments we thought of the following features that were significant in our analysis.

1. **Average Out Of Pocket Payment:** Generally Out of Pocket medical expenses include deductibles, coinsurance, and copayments for the covered services. Out of Pocket payments also include expenses which are not covered by the medical



insurance.

We generated out of pocket expenses by taking the difference of average total payments and average medicare payments.

2. **Average Extra Payment:** This feature represented extra charges imposed by the hospital in a particular state for a specific DRG code. In the medicare industry over billing is always a concern when it comes to analysing payments and performance. Hence keeping this in mind we thought of generating this feature which helped us to calculate extra cost billed by the provider. Average Extra Payments included hospital charges for Beds, Room facilities, Preliminary tests, Diet Plans, Medical Tests and Other amenities which are not covered by the medicare.

We formulated this feature by taking the difference of average covered charges and average total payments.

3. **Percentage Payment Reduction:** All the facilities in the United States are evaluated based on four parameters and are assigned performance scored. The four criterias are as below:
 1. Clinical Outcomes Score
 2. Efficiency and cost reduction score
 3. Safety Domain Score
 4. Person and Community Engagement Score.

If a hospital's total performance score lags behind the national average score, that is 61 then on 10 points lag the hospitals receive 0.5% reduction in the medicare payment. In this project we worked on the data which had more than 2500 unique hospitals. Hence we thought it might be interesting to see the payment reduction percentage for each facility based on its performance score.

7. Predictor/Variable Choice

The table below shows the relevant variables used for the analysis along with their hypothetical impact and rationale

Variables	Effect	Explanation
Total Discharges	+/-	Hospitals with more discharges per DRG might have optimized for efficiency & better outcomes leading to increased payments
Average Covered Charges	+	Represents the provider's billing for services covered by Medicare. Since these vary across hospitals, providers with higher billing might receive more payment from Medicare.
Clinical Outcomes Score	+	Measures mortality and patient's improvement in condition for Heart Failure, Myocardial Infarction and Pneumonia against benchmark and baseline levels. A hospital with a greater score is less likely to receive cuts in Medicare payments.
Community & Engagement Score	+	Measures hospital performance based on communication with nurses, doctors, discharge information, hospital cleanliness & quietness. Higher scores → lesser payment cuts from Medicare.
Efficiency & Cost Reduction Score	+/-	It is a function of medicare spending per beneficiary. Higher scores result in lesser payment cuts. However, this could be influenced by population, proportion of low income patients and wage levels.
Safety Score	+	Measures hospital's efforts to prevent infections such as those occurring at Surgical Site, Catheter Associated UTI, Central Line Associated bloodstream infection being transmitted to patients. Higher scores → no payment reduction from Medicare.
DRG Code	+/-	Each procedure / medical condition is codified as a Diagnosis Related Group (DRG).
CCN / Facility ID	Random Effect	CMS Certification Number for each individual medical facility in the data set. Can be used to assess variations for a particular DRG across facilities.
State	Random	The state in which the medical facility is located. Can be

	m Effect	used to measure the variations between states across the DRG groups.
Payment Reduction Percentage	-	Measure hospital performance score vs. national average. For every 10 points drop in score below the national avg., Medicare withholds 0.5% of the payouts to be made to the facility.
Average_Medicare_Payment	+/-	Average Medicare payments can be crucial in analysing average out of pocket payment as higher medicare might affect deductibles and copays positively or negatively.

8. Model Building and Interpretations.

Looking at the dataset first thing which we identified is that our dataset is a multi level dataset as it contains drg codes, facility ID along with the location. Hence we considered DRG Codes as level 1, Facility IDs as level 2 and Provider State as level 3 in our analysis. According to our research we thought of the most relevant and important three dependent variables which generated business value and targeted vital aspects of the IPPS. Three Dependent Variables.

- 1) Average Medicare Payments
- 2) Average Out of Pocket Payments
- 3) Average Extra Charges

Considering these three variables we answered the following questions with our analysis.

1) How do out-of-pocket charges vary across the United State for each Diagnosis Related Group (DRG)?

Model 1:

```
Avg_outofpocket_pymnts_states =  
lmer(avg_outofpocket_pymnts ~ total_discharges + avg_covered_charges +  
avg_medicare_pymnts + weighted_clinical_score + weighted_community_score +  
weighted_costreduction_score + weighted_safety_score + ( 1 | provider_state/drg_code ), data =  
clean_dataset, REML = FALSE)
```

Interpreting Fixed Effect Variables

Total Discharges	1 unit increase in total discharges will increase the average out of pocket payment by \$77.35
Average Covered Charges	\$1 increase in average covered charges will increase the average out of pocket payments by \$740.685
Average Medicare Payments	\$1 increase in average medicare payments will increase the average out of pocket payments by \$891.803
Weighted Normalized Clinical Outcomes Domain	1 unit increase in weighted clinical scores will increase the average out of pocket payments by \$47.878

Score	
Weighted Person and Community Engagement Domain Score	1 unit increase in weighted community scores will increase the average out of pocket payments by \$131.230
Weighted Efficiency and Cost Reduction Domain Score	1 unit increase in weighted cost reduction scores will decrease the average out of pocket payments by \$98.126
Weighted Safety Domain Score	1 unit increase in weighted safety scores will decrease the average out of pocket payments by \$124.692

Interpreting Random Effect Variables

Top 3 DRG codes with highest out of pocket expense.

DRG Code	DRG Description	Average payments above the mean.
008	SIMULTANEOUS PANCREAS/KIDNEY TRANSPLANT	22335.80
005	LIVER TRANSPLANT W MCC OR INTESTINAL TRANSPLANT	18743.08
007	LUNG TRANSPLANT	15595.09

State	Average payments above the mean.
Hawaii	2094.55

Utah	613.33
Virginia	585.13

2) How do extra charges (service charges) vary across different medical facilities in the state of Florida?

Model 2:

avg_extra_pymnts_fl =

```
lmer (avg_extra_pymnts ~ total_discharges + avg_covered_charges + avg_medicare_pymnts +
weighted_clinical_score + weighted_community_score + weighted_costreduction_score +
weighted_safety_score + ( 1 | facility_id ) + ( 1 | drg_code ), data = subset (clean_dataset,
provider_state == "FL" ), REML = FALSE)
```

Interpreting Fixed Effect Features/Variables

Total Discharges	1 unit increase in total discharges will decrease the average extra payment by \$30.67
Average Covered Charges	\$1 increase in average covered charges will increase the average extra payments by \$66298.25
Average Medicare Payments	\$1 increase in average medicare payments will decrease the average extra payments by \$13.835
Weighted Normalized Clinical Outcomes Domain Score	1 unit increase in weighted clinical scores will decrease the average extra payments by \$38.61
Weighted Person and Community Engagement Domain Score	1 unit increase in weighted community scores will decrease the average extra payments by \$238.83
Weighted Efficiency	1 unit increase in weighted cost reduction scores will decrease the average extra payments by \$71.47

and Cost Reduction Domain Score	
Weighted Safety Domain Score	1 unit increase in weighted safety scores will decrease the average extra payments by \$124.692

Interpreting Random Effect Variables

Top 3 Facilities which charge maximum extra charges in Florida

Facility ID	Facility Name	Extra charges higher than the mean in \$ (Sorted by Most Expensive)
100038	MEMORIAL REGIONAL HOSPITAL	907.47
100315	VIERA HOSPITAL	889.97
100183	CORAL GABLES HOSPITAL	675.37

Top 3 DRG codes for which facilities charge maximum extra charges in Florida

DRG	DRG Description	Extra charges higher than the mean in \$ (Sorted by Most Expensive)
003	ECMO OR TRACH W MV >96 HRS OR PDX EXC FACE, MOUTH & NECK W MAJ O.R.	5898.48
266	ENDOVASCULAR CARDIAC VALVE REPLACEMENT W MCC	5037.69
456	SPINAL FUS EXC CERV W SPINAL CURV/MALIG/INFEC OR EXT FUS W MCC	4507.38

3) Are the medical providers paid the same across the United States by the Medicare for each Diagnosis Related Group (DRG)?

Model 3:

avg_medicare_payments =

```
lmer (avg_medicare_payments ~ as.factor(drg.code) + total_score + total_discharges + payment_reduction + avg_income + population + (1 | state), data = d, REML = FALSE)
```

While it is expected that Medicare payouts account for differences in location, we will be interpreting the random effects to understand which three states have the highest and lowest variance in avg. medicare payouts.

HIGHEST PAYOUTS

State	Avg. Medicare Payout (across all DRGs)	Prevailing Wage Levels (for reference)
New York	+ \$12,634	\$61,543
Oregon	+ \$12,030	\$62,498
Hawaii	+ \$8,662	\$73,599

LOWEST PAYOUTS

State	Avg. Medicare Payout (across all DRGs)	Prevailing Wage Levels (for reference)
Virginia	- \$7,513	\$70,811
Tennessee	- \$6,373	\$55,306
Iowa	- \$4,968	\$63,467

9. Recommendations & Future Scope

1. Hospitals in the states of Hawaii, Utah, and Virginia should probe into their billing as they seem to have the highest out of pocket expenses, which has a higher chance of not being repaid since it is paid off by the patients.
2. Medicare should probe into hospitals such as the Memorial Regional Hospital, Viera Hospital and Coral Gables Hospitals as they are among the highest in terms of extra charges billed to patients resulting in expensive medical treatment bills for the DRG's as compared to other facilities.
3. State of Hawaii has the maximum medicare payment considering other states having similar average household income and population.
4. State of Virginia has the minimum medicare payment considering other states having similar average household income and population.
5. Hospitals should focus on improving their Total Performance Score in order to minimize losses in terms of payment cuts from Medicare. Out of the 2500 unique medical facilities in our dataset, only 105 providers had a TPS ≥ 59 (the threshold beyond which payment reductions kick in).

Considering the future scope of our project we are looking forward to blend datasets based on hospitalizations, physicians associated with specific hospitals and perform analysis in more detail. We would like to extend our project by considering data specific to Hawaii, New York, and Oregon state as they have maximum medicare payments. Similarly we are looking forward to diving deep into the states of Hawaii, Utah and Virginia and perform regression analysis in more detail oriented fashion. In a nutshell, our future goal for this project is to drill one level deep into medicare IPPS payments and merging hospital specific, standard of living and physician data for performing statistical data analysis.

10. References

- <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Information-on-Prescription-Drugs/MedicarePartD>
- https://www.cms.gov/outreach-and-education/medicare-learning-network-mln/mln-products/downloads/hospital_vbpurchasing_fact_sheet_icn907664.pdf
- <https://www.medicare.gov/hospitalcompare/data/total-performance-scores.html>
- <https://rpubs.com/CoData/MedicareFeeForService>
- <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HVBP/Hospital-Value-Based-Purchasing>

Appendix. R code for the project

```
rm(list=ls())
#-----IMPORTING LIBRARIES-----

library(readxl)
library(plyr)
library(tidyr)
library(lme4)
library(dplyr)

#-----IMPORTING THE DATASETS-----
#-----IMPORTING HVP PERFORMANCE SCORES-----

perf_score = read_excel("performance_scores.xlsx")
colnames(perf_score) = tolower(make.names(colnames(perf_score)))
attach(perf_score)
summary(perf_score)

#-----IPPS FY 2017 DATA-----

ipps_payment = read_excel("IPPS FY 2017 Data.xlsx")
colnames(ipps_payment) =
tolower(make.names(colnames(ipps_payment)))
attach(ipps_payment)
summary(ipps_payment)

#-----Statewise Population-----
population = read_excel("State County Level Population
2017.xlsx")
colnames(population) = tolower(make.names(colnames(population)))
attach(population)
summary(population)

#----- State Wise Average Income -----

income = read_excel("house_income.xlsx")
colnames(income) = tolower(make.names(colnames(income)))
attach(income)
summary(income)

#---RENAMING VARIABLE/FEATURES TO MAKE THEM CONSISTENT---
```

```

names(ipps_payment)
names(perf_score)

names(perf_score)[names(perf_score) == "?..facility.id"] <-
"facility.id"
names(ipps_payment)[names(ipps_payment) == "provider.id"] <-
"facility.id"

#---PERFORMING INNER JOIN TO COMBINE BOTH DATASETS BASED ON
'FACILITY_ID'-----

master_dataset = join(ipps_payment, perf_score, type = "inner")
# Join should be by facility ID
income_population = join(population, income, type = "inner")
#Merging Income and Population Dataset
#master_dataset = join(master_dataset, income_population, type =
"inner") #Final dataset

#-----DATA PREPROCESSING-----

names(master_dataset)
master_dataset = separate(data = master_dataset, col =
drg.definition, into = c("drg.code", "drg.description"), sep =
"\\-")
master_dataset = separate(data = master_dataset, col =
hospital.referral.region..hrr..description, into =
c("hospital.state", "hospital.region"), sep = "\\-")

#-----IMPORTING FINAL PREPROCESSED DATA INTO A NEW
DATAFRAME-----

clean_dataset = subset(master_dataset, select =
-c(facility.name, address, city, state, zip.code, location,
hospital.state, hospital.region))

clean_dataset <-
clean_dataset[!(clean_dataset$unweighted.normalized.clinical.out
comes.domain.score == "Not Available"
|
clean_dataset$weighted.normalized.clinical.outcomes.domain.score
== "Not Available"
|
clean_dataset$unweighted.person.and.community.engagement.domain.
score == "Not Available")

```

```

|
clean_dataset$weighted.person.and.community.engagement.domain.score == "Not Available"
|
clean_dataset$unweighted.normalized.safety.domain.score == "Not Available"
|
clean_dataset$weighted.safety.domain.score == "Not Available"
|
clean_dataset$unweighted.normalized.efficiency.and.cost.reduction.domain.score == "Not Available"
|
clean_dataset$weighted.efficiency.and.cost.reduction.domain.score == "Not Available"
),]

```

#-----Changing Columns names-----

```

names(clean_dataset)[names(clean_dataset) == "drg.code"] <- "drg_code"
names(clean_dataset)[names(clean_dataset) == "drg.description"] <- "drg_description"
names(clean_dataset)[names(clean_dataset) == "facility.id"] <- "facility_id"
names(clean_dataset)[names(clean_dataset) == "provider.name"] <- "provider_name"
names(clean_dataset)[names(clean_dataset) == "provider.street.address"] <- "provider_street_address"
names(clean_dataset)[names(clean_dataset) == "provider.city"] <- "provider_city"
names(clean_dataset)[names(clean_dataset) == "provider.zip.code"] <- "provider_zipcode"
names(clean_dataset)[names(clean_dataset) == "provider.state"] <- "provider_state"
names(clean_dataset)[names(clean_dataset) == "total.discharges"] <- "total_discharges"
names(clean_dataset)[names(clean_dataset) == "average.covered.charges"] <- "avg_covered_charges"
names(clean_dataset)[names(clean_dataset) == "average.total.payments"] <- "avg_total_pymnts"
names(clean_dataset)[names(clean_dataset) == "average.medicare.payments"] <- "avg_medicare_pymnts"
names(clean_dataset)[names(clean_dataset) == "county.name"] <-

```

```

"county"
names(clean_dataset)[names(clean_dataset) ==
"unweighted.normalized.clinical.outcomes.domain.score"] <-
"unweighted_clinical_score"
names(clean_dataset)[names(clean_dataset) ==
"weighted.normalized.clinical.outcomes.domain.score"] <-
"weighted_clinical_score"
names(clean_dataset)[names(clean_dataset) ==
"unweighted.person.and.community.engagement.domain.score"] <-
"unweighted_community_score"
names(clean_dataset)[names(clean_dataset) ==
"weighted.person.and.community.engagement.domain.score"] <-
"weighted_community_score"
names(clean_dataset)[names(clean_dataset) ==
"unweighted.normalized.safety.domain.score"] <-
"unweighted_safety_score"
names(clean_dataset)[names(clean_dataset) ==
"weighted.safety.domain.score"] <- "weighted_safety_score"
names(clean_dataset)[names(clean_dataset) ==
"unweighted.normalized.efficiency.and.cost.reduction.domain.score"] <- "unweighted_costreduction_score"
names(clean_dataset)[names(clean_dataset) ==
"weighted.efficiency.and.cost.reduction.domain.score"] <-
"weighted_costreduction_score"
names(clean_dataset)[names(clean_dataset) ==
"total.performance.score"] <- "total_performance_score"
names(clean_dataset)[names(clean_dataset) ==
"average.income.per.household"] <- "avg_income/house"

#----- Generating Features -----

clean_dataset$avg_outofpocket_pymnts =
(clean_dataset$avg_total_pymnts -
clean_dataset$avg_medicare_pymnts)

clean_dataset$avg_extra_pymnts =
clean_dataset$avg_covered_charges -
clean_dataset$avg_total_pymnts

#-----converting scores to numeric and removing outlier
values-----

clean_dataset$unweighted_clinical_score =
as.numeric(clean_dataset$unweighted_clinical_score)

```



```

clean_dataset$weighted_clinical_score =
as.numeric(clean_dataset$weighted_clinical_score)
clean_dataset$unweighted_community_score =
as.numeric(clean_dataset$unweighted_community_score)
clean_dataset$weighted_community_score =
as.numeric(clean_dataset$weighted_community_score)
clean_dataset$unweighted_safety_score =
as.numeric(clean_dataset$unweighted_safety_score)
clean_dataset$weighted_safety_score =
as.numeric(clean_dataset$weighted_safety_score)
clean_dataset$unweighted_costreduction_score =
as.numeric(clean_dataset$unweighted_costreduction_score)
clean_dataset$weighted_costreduction_score =
as.numeric(clean_dataset$weighted_costreduction_score)
is.num <- sapply(clean_dataset, is.numeric)
is.num
clean_dataset[is.num] <- lapply(clean_dataset[is.num], round, 4)

#-----Some Visualization-----
library(lattice)
bwplot(~avg_outofpocket_pymnts | provider_state,
data=clean_dataset,
      main = "Out of Pocket Payments")

bwplot(~avg_covered_charges | provider_state,
data=clean_dataset,
      main = "Covered Charges")

#-----Scaling variables-----
clean_dataset$total_discharges =
scale(clean_dataset$total_discharges)
clean_dataset$avg_covered_charges =
scale(clean_dataset$avg_covered_charges)
clean_dataset$avg_total_pymnts =
scale(clean_dataset$avg_total_pymnts)
clean_dataset$avg_medicare_pymnts =
scale(clean_dataset$avg_medicare_pymnts)

clean_dataset$weighted_clinical_score =
scale(clean_dataset$weighted_clinical_score)
clean_dataset$weighted_community_score =
scale(clean_dataset$weighted_community_score)

```

```

clean_dataset$weighted_safety_score =
scale(clean_dataset$weighted_safety_score)
clean_dataset$weighted_costreduction_score =
scale(clean_dataset$weighted_costreduction_score)

#clean_dataset$avg_outofpocket_pymnts =
scale(clean_dataset$avg_outofpocket_pymnts)
#clean_dataset$avg_extra_pymnts =
scale(clean_dataset$avg_extra_pymnts)

colSums(is.na(clean_dataset))
sum(is.na(clean_dataset))

attach(clean_dataset)
#----storing updated dataframe into the file for reference-----
#write.csv(clean_dataset, "project_dataset.csv")
#dataset = read.csv("project_combined_dataset_final.csv")

#-----DATA EXPLORATION-----

#-----EXPLORING THE DISTRIBUTION OF THE PREDICTORS-----

hist(total_discharges, col="skyblue") #Rightskewed
hist(log(total_discharges), col="green") #Looks better hence
could be used as the predictor.

hist(avg_outofpocket_pymnts, col = "blue")
hist(log(avg_outofpocket_pymnts), col = "pink")

hist(avg_extra_pymnts, col = "blue")
hist(log(avg_extra_pymnts), col = "pink")

#----CONVERTING CATEGORICAL VARIABLES INTO FACTORS----
clean_dataset$drg_code <- factor(clean_dataset$drg_code)
levels(clean_dataset$drg_code)

clean_dataset$provider_city <-
factor(clean_dataset$provider_city)
levels(clean_dataset$provider_city)

clean_dataset$provider_state <-
factor(clean_dataset$provider_state)
levels(clean_dataset$provider_state)

```

```

clean_dataset$county <- factor(clean_dataset$county)
levels(clean_dataset$county)

clean_dataset$facility_id <- factor(clean_dataset$facility_id)
levels(clean_dataset$facility_id)

#----- Random Effect Model Average Payment varying
across DRG codes in Different states -----

#4) How much Out of pocket expense vary according to state?
avg_outofpocket_pymnts_states =
  lmer(avg_outofpocket_pymnts ~ total_discharges +
avg_covered_charges
                                + avg_medicare_pymnts +
weighted_clinical_score
                                + weighted_community_score +
weighted_costreduction_score
                                + weighted_safety_score + (1
| provider_state/drg_code),
                                data = clean_dataset, REML =
FALSE)
summary(avg_outofpocket_pymnts_states)

confint(avg_outofpocket_pymnts_states)
AIC(avg_outofpocket_pymnts_states)
fixef(avg_outofpocket_pymnts_states)      # Magnitude of
fixed effect
ranef(avg_outofpocket_pymnts_states)      # Magnitude of
random effect
coef(avg_outofpocket_pymnts_states)

fl_d = subset(clean_dataset, provider_state == "FL")

avg_extra_pymnts_fl =
  lmer(avg_extra_pymnts ~ total_discharges + avg_covered_charges
+
  + avg_medicare_pymnts + weighted_clinical_score
  + weighted_community_score + weighted_costreduction_score
  + weighted_safety_score + (1 | facility_id) + (1 |
drg_code),
  data = subset(clean_dataset, provider_state == "FL") , REML
= FALSE)
summary(avg_extra_pymnts_fl)
ranef(avg_extra_pymnts_fl)

```

```

options(max.print = 100000000)

unique(clean_dataset$facility_id)
unique(fl_d$facility_id)

unique(clean_dataset$drg_code)
unique(fl_d$drg_code)

l1 = lmer(avg_medicare_payments ~ drg.code +
weighted_clinical_outcomes + weighted_engagement_community_score
+ weighted_safety_score + weighted_efficiency_score +
avg_covered_charges + (1 | state), data = final, REML
= FALSE)

final$drg.code =
as.numeric(levels(final$drg.code))[final$drg.code]

top_drgs = subset(final, final$drg.code == 853 | final$drg.code
== 003 | final$drg.code == 870 | final$drg.code == 329)
m2 = lmer(avg_medicare_payments ~ as.factor(drg.code) +
total_score + total_discharges + payment_reduction +
(1 | state), data = top_drgs, REML = FALSE)

summary(m2)
ranef(m2)
confint(m2)
AIC(m2)

m3 = lmer(avg_medicare_payments ~ as.factor(drg.code) +
total_score + total_discharges +
payment_reduction + avg_income + population + (1 |
state), data = top_drgs, REML = FALSE)
summary(m3)
ranef(m3)
plot(m3)

```